

II. Population Genetics

Haplotype Inference

Lecture 17,18

Instructors: Su-In Lee & Ming-Chi Tasi

Last Lecture

- Introduction to the E-M algorithm
- Approach for likelihood optimization
- Examples related to gene counting
 - Allele frequencies estimation
 - Haplotype frequency estimation

Today's Lecture

- Other approaches for haplotyping
 - Clark's *greedy* algorithm
 - Stephens et al's coalescent based algorithm
- Using haplotypes in association studies

Useful Roles for Haplotypes

- Linkage disequilibrium studies
 - Summarize genetic variation
- Selecting markers to genotype
 - Identify haplotype tag SNPs
- Candidate gene association studies
 - Help interpret single marker associations
 - May capture effect of ungenotyped alleles

The problem

- Haplotypes are hard to measure directly
 - X-chromosome in males
 - Sperm typing
 - Hybrid cell lines
 - Other molecular techniques
- Often, statistical reconstruction required

Typical Genotype Data

- Two alleles for each individual
 - Chromosome origin for each allele is unknown

Observation

C	G	Marker1
T	C	Marker2
G	A	Marker3

Possible States

C	G	C	G
T	C	C	T
G	A	G	A
C	G	C	G
C	T	T	C
A	G	A	G

- Multiple haplotype pairs can fit observed genotype

Use Information on Relatives?

- Family information can help determine phase at many markers
- Still, many ambiguities might not be resolved
 - Problem more serious with larger numbers of markers
- Can you propose examples?

What if there are no relatives?

- Rely on linkage disequilibrium
- Assume that population consists of small number of distinct haplotypes
- Haplotypes tend to be similar

Clark's Haplotyping Algorithm

- Clark (1990) *Mol Biol Evol* 7:111-122
- One of the first haplotyping algorithms
 - Computationally efficient
 - Very fast and widely used in 1990's
 - More accurate methods are now available

Clark's Haplotyping Algorithm

- Find unambiguous individuals
 - What kinds of genotypes will these have?
 - Initialize a list of known haplotypes
- Resolve ambiguous individuals
 - If possible, use two haplotypes from list
 - Otherwise, use one known haplotype and augment list
- If unphased individuals remain
 - Assign phase randomly to one individual
 - Augment haplotype list and continue from previous step

Parsimonious Phasing - Example

1 0 1 0 0 h

h 0 1 h 0 0

0 h h 1 h 0

1 0 1 0 0 0
1 0 1 0 0 1

1 0 1 0 0 0
0 0 1 1 0 0

0 0 1 1 0 0
0 1 0 1 1 0

Notes ...

- Clark's Algorithm is extremely fast
- More likely to start with large sample
- Orphaned alleles and anomalous matches may occur
 - Solution with the least orphaned alleles is usually the one with the fewest anomalous matches

The EM Haplotyping Algorithm

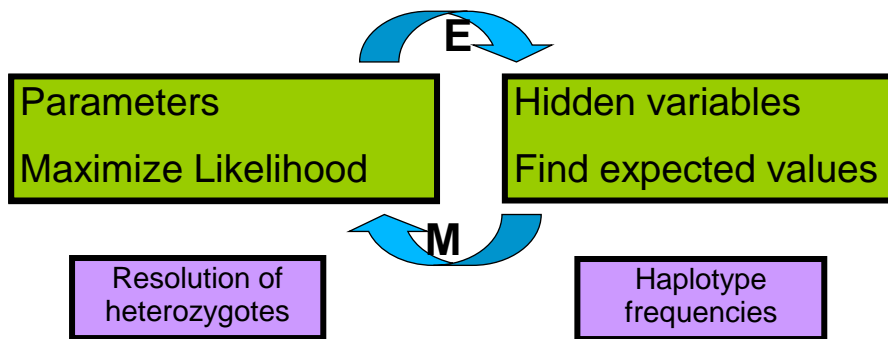
- Excoffier and Slatkin (1995)
 - *Mol Biol Evol* **12**:921-927
 - Provide a clear outline of how the algorithm can be applied to genetic data
- Combination of two strategies
 - E-M statistical algorithm for missing data
 - Counting algorithm for allele frequencies

EM Algorithm For Haplotyping

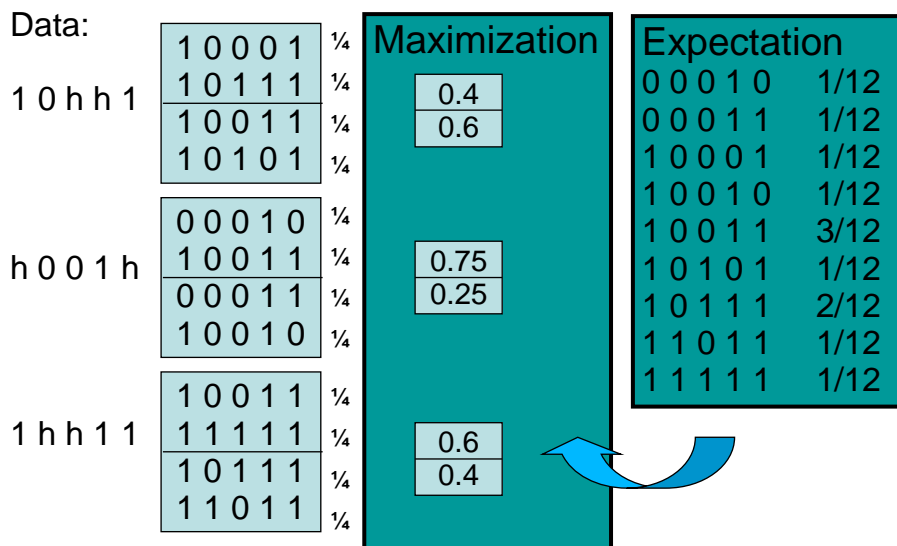
1. “Guesstimate” haplotype frequencies
2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
3. Estimate frequency of each haplotype by counting
4. Repeat steps 2 and 3 until frequencies are stable

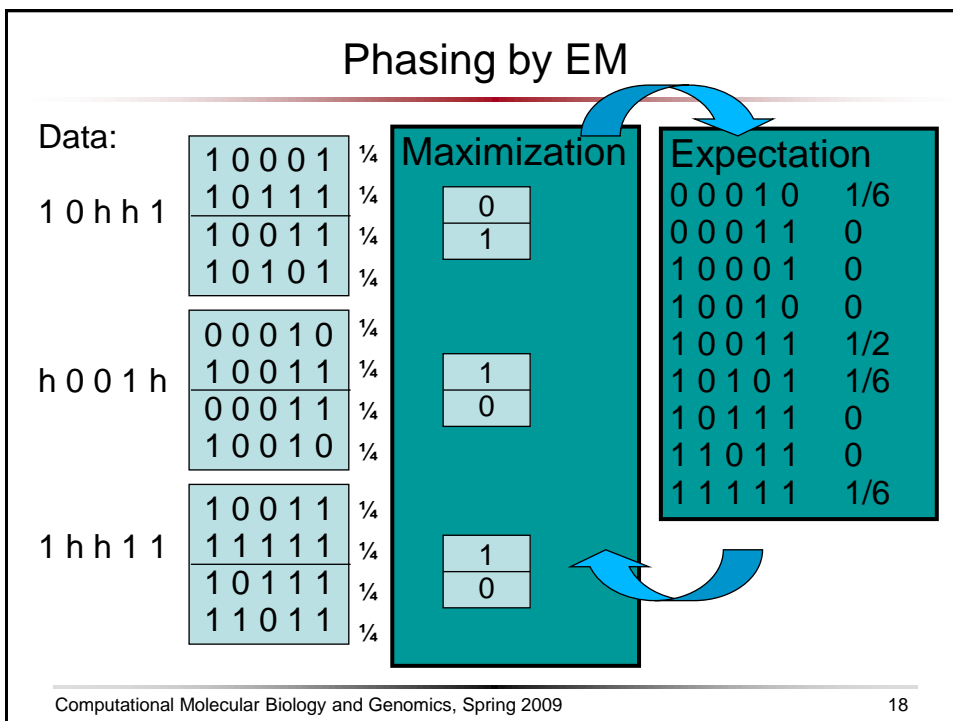
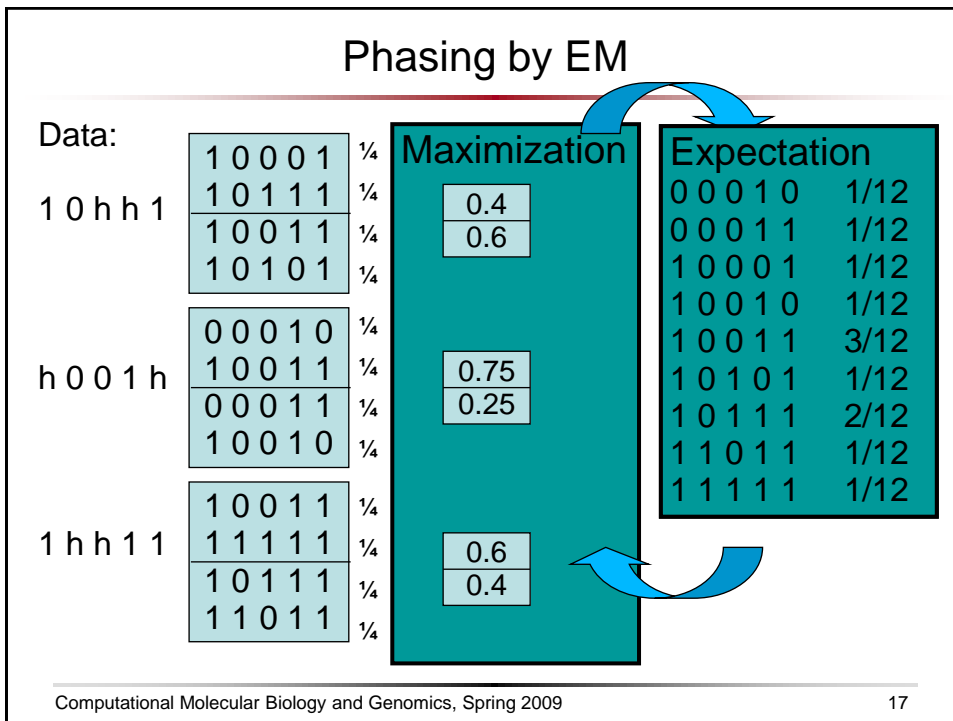
Phasing by EM

- E-M:
 - Method for maximum-likelihood parameter inference with hidden variables



Phasing by E-M





Expected Haplotype Counts

h_j = haplotype j

$G(h_i, h_j)$ = Unphased genotype corresponding to h_i, h_j

n_G = Number of genotypes of type G

$H \sim G$ = Haplotype pairs compatible with G

$$E(n_{h_i}) = 2n_{G(h_i, h_i)} + \sum_{h_j \neq h_i} n_{G(h_i, h_j)} \frac{2\hat{p}_{h_i}\hat{p}_{h_j}}{\sum_{H \sim G(h_i, h_j)} P(H | \hat{p})}$$

Computational Cost (for SNPs)

- Consider sets of m unphased genotypes
 - Markers 1.. m
- If markers are bi-allelic
 - 2^m possible haplotypes
 - $2^{m-1} (2^m + 1)$ possible haplotype pairs
 - 3^m distinct observed genotypes
 - 2^{n-1} reconstructions for n heterozygous loci
- For example, if $m = 10$

EM Algorithm For Haplotyping

- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
 - Fewer microsatellites
- More accurate than Clark's method
- Fully or partially phased individuals contribute most of the information

Enhancements to EM

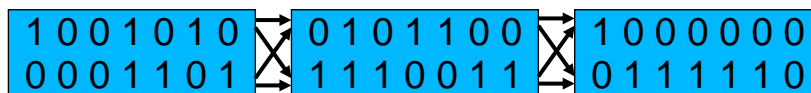
- List only haplotypes present in sample
- Gradually expand subset of markers under consideration, eliminating haplotypes with low estimated frequency from consideration at each stage
 - SNPHAP [Clayton (2001)]
 - HAPLOTYPER [Qin et al. (2002)]

Divide-And-Conquer Approximation

- Number of potential haplotypes increases exponentially
 - Number of observed haplotypes does not
- Approximation
 - Successively divide marker set
 - Run E-M on each segment
 - Prune haplotype list as segments are ligated
- Computation order: $\sim m \log m$
 - Exact EM is order $\sim 2^m$

Partition-Ligation EM

- In practice:
 - #variables is exponential in too many sites
- Solution:
 - Locally phase each region
 - Merge by phasing vectors of haplotype pairs



Other Recent Developments ...

- Newer methods try to further improve haplotype estimation by favoring sets of similar haplotypes
- Stephens et al. (2001)
 - *Am J Hum Genet* **68**:978-89
- Genealogical approach...

What the Genealogy Implies ...

- Haplotypes are similar to each other...

Known Haplotypes

22544
22544
22544

33334
33334

23233

14234

Individual 1

Genotype:

32344
23534

Individual 2:

Genotype:

32444
23434

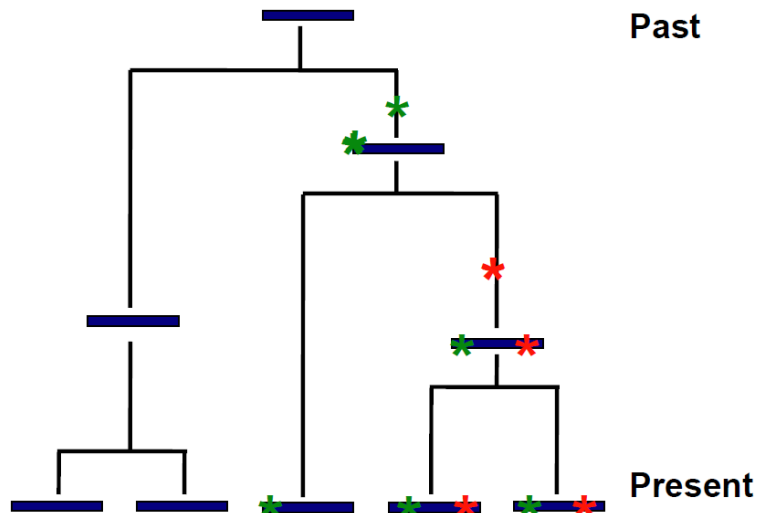
33334

22544

33434

22444

Chromosome Genealogies



Method based on Gibbs sampler

- MCMC method
 - Stochastic, random procedure
 - Improves solution gradually
- Given initial set of haplotypes
- Sample haplotypes for one individual at a time, assuming other haplotypes are true
- Repeat a few million times...

Update Procedure I

- Pick individual U to update at random
- Calculate haplotype frequencies F in all other individuals
 - Since everyone is “phased”, this is done by counting
- Sample new haplotypes for U from conditional distribution of U 's haplotypes given F

Update Procedure I

- This procedure would produce an estimate of haplotype frequencies that equivalent to the E-M algorithm...
- Stephens et al (2001) suggested an alternative estimate of F ...

Update Procedure II

- Estimate F from the other individuals
- Construct F^* to include haplotypes in F and also other similar (possibly differing at a few sites)
- Update U 's haplotypes conditional on F^*

Stephens' Formula ...

- $\Pr(h|H)$ is the probability of observing haplotype h given previous set H

$$\Pr(h | H) = \sum_{\alpha} \sum_s \frac{n_{\alpha}}{n} \left(\frac{\theta}{n + \theta} \right)^s \frac{n}{n + \theta} (P^s)_{\alpha h}$$

Sum over haplotypes Sum over number of mutations S mutations before coalescence Coalescence Mutation Matrix

Further Refinements

- This naïve strategy becomes impractical for very long haplotypes
 - List of haplotypes for each individual could become too long
- Instead, we can proceed by selecting a short segment of the haplotype to update at random

Hypothesis Testing

- Often, haplotype frequencies are not final outcome.
- For example, we may wish to compare two groups of individuals...
 - Are haplotypes similar in two populations?
 - Are haplotypes similar in patients and healthy controls?

Acknowledgement

- This lecture note is based on material by Profs Goncalo Abecasis (Univ of Michigan) and Itsik Pe'er (Columbia Univ).